PATENT ABSTRACTS OF JAPAN

(11)Publication number : **10-269248**
(43)Date of publication of application : **09.10.1998**

(51)Int.Cl. **G06F 17/30**

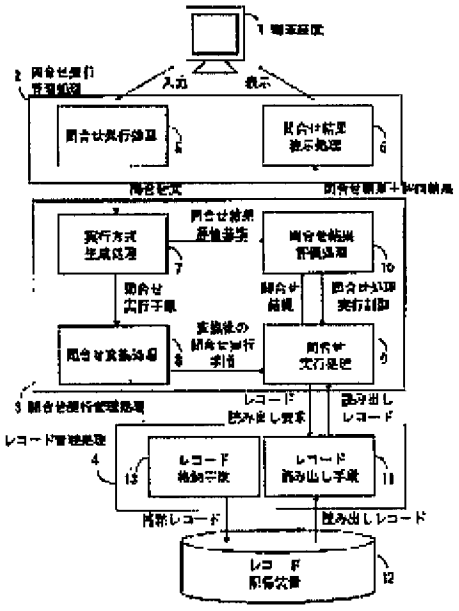
(21)Application number : **10-026493**
(22)Date of filing : **23.01.1998**

(71)Applicant : **HITACHI LTD**
(72)Inventor : **USHIJIMA KAZUTOMO**
FUJIWARA SHINJI
MASAI KAZUO
TAKAHASHI YORI
NISHIZAWA ITARU

(30)Priority
Priority number : **09 25863** ????Priority date : **24.01.1997** ????Priority country : **JP**

(54) METHOD FOR RANDOM EXTRACTION OF DATA IN DATA BASE PROCESSING SYSTEM, AND DATA BASE PROCESSING SYSTEM BASED UPON THE SAME

(57)Abstract:
PROBLEM TO BE SOLVED: To improve the throughput of the random extracting processing in a data base processing system.
SOLUTION: It is made possible to issue an inquiry including random extraction in an inquiry issuing process 2, and application order between the random extraction and other inquiries is changed in an inquiry converting process 8 in consideration of extraction units of the random extraction. Further, random access to a secondary storage device is reduced in a record managing process 4. Thus, the inquiry including the random extraction can be issued and inquiry conversion is performed in consideration of the extraction units to enable application to an inquiry including a totalizing process and further improve the efficiency of inquiries over a wide range. The random access to the secondary storage device is reduced to be able to improve the efficiency more.



LEGAL STATUS

[Date of request for examination] 02.03.2004

[Date of sending the examiner s decision of rejection]

[Kind of final disposal of application other than the examiner s decision of rejection or application converted registration]

[Date of final disposal for application]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平10－269248

(43)公開日　平成10年(1998)10月９日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/401	３２０Ｚ
			15/40	３８０Ｄ
			15/403	３４０Ｄ
				３７０Ｚ

審査請求　未請求　請求項の数16　ＦＤ　（全 21 頁）

(21)出願番号　特願平10－26493

(22)出願日　平成10年(1998)１月23日

(31)優先権主張番号　特願平9－25863

(32)優先日　平９(1997)１月24日

(33)優先権主張国　日本（ＪＰ）

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　牛嶋　一智
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内

(72)発明者　藤原　眞二
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内

(72)発明者　正井　一夫
神奈川県横浜市戸塚区戸塚町5030番地　株式会社日立製作所ソフトウェア開発本部内

(74)代理人　弁理士　笹岡　茂　（外１名）

最終頁に続く

(54)【発明の名称】　データベース処理システムにおけるデータの無作為抽出処理方法及びこれに基づくデータベース処理システム

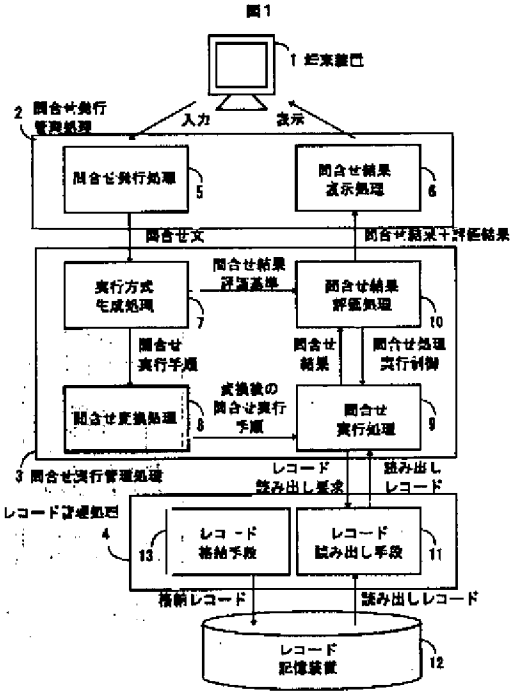
(57)【要約】

【課題】　データベース処理システムにおける無作為抽出処理の処理効率向上を図るための問合せ発行処理方法及び問合せ変換処理方法及びデータ管理処理方法を提供する。

【解決手段】　問合せ発行処理２にて無作為抽出処理を含む問合せの発行を可能にし問合せ変換処理８において無作為抽出処理と他の問合せとの間の適用順序を無作為抽出処理の抽出単位を考慮しながら交換する。さらにレコード管理処理４において二次記憶装置に対するランダムアクセスを削減する。

【効果】　無作為抽出処理を含む問合せの発行を可能にし、抽出単位を考慮した問合せ変換を行うことで集計処理を含む問合せに対しても適用でき、さらに広い範囲の問合せの効率向上を図れる。二次記憶装置に対するランダムアクセスを削減することで一層の効率向上を図れる。

図1



【特許請求の範囲】

【請求項１】　データベースより所望のデータを抽出するデータベース処理システムにおける無作為抽出処理方法であって、（１）前記データベースに対する問合せを発行する問合せ発行管理処理と、（２）発行された問合せの実行管理を行う問合せ実行管理処理と、（３）前記データベースに対するデータの格納及び格納されたデータの管理を行うデータ管理処理とを有し、前記（２）の問合せ実行管理処理は、（２-１）前記問合せ発行処理で発行された問合せに無作為抽出処理を挿入する処理と、（２-２）挿入された無作為抽出処理における抽出単位を保存しながら前記問合せをさらに高効率なデータ抽出を行い得る問合せに変換する問合せ変換処理と、を有することを特徴とするデータベース処理システムにおける無作為抽出処理方法。

【請求項２】　前記（２）の問合せ実行管理処理で実行される問合せ処理が、（Ａ）処理対象のデータとしてのレコードをテーブル化した表に対して、前記（１）で発行された問合せで指定された一つあるいは複数のグループ化カラムの値に応じて表のレコードをグループに分け、同じく前記（１）で発行された問合せで指定された一つあるいは複数の集計カラムの値に関してそれぞれのグループ毎に集計を行う分類集計処理を適用した後、（Ｂ）問合せで指定された一つあるいは複数の無作為抽出カラムのそれぞれに無作為に値を指定し、指定された値を持つレコードを前記集計結果から全て抽出する無作為抽出処理を適用する、からなる処理を含む場合、前記（２-２）の問合せ変換処理は、前記表のレコードに対して、処理（Ｂ）を適用した後、処理（Ａ）を適用するように、該処理（Ａ）と（Ｂ）の適用順序を交換する問合せ変換処理を備える請求項１記載のデータベース処理システムにおける無作為抽出処理方法。

【請求項３】　前記（２－２）の問合せ変換処理は、前記処理（Ｂ）が無作為抽出カラムを指定しないレコード単位の無作為抽出処理である場合には、前記無作為抽出カラムとして前記処理（Ａ）の分類集計処理のグループ化カラムを用いる問合せに変換した後で実行される請求項２記載のデータベース処理システムにおける無作為抽出処理方法。

【請求項４】　前記（２）の問合せ実行管理処理で実行される問合せ処理が、（Ａ）処理対象のデータとしてのレコードをテーブル化した表に対して、前記（１）で発行された問合せで指定された一つあるいは複数のグループ化カラムの値に応じて表のレコードをグループに分け、（Ｂ）同じく前記（１）で発行された問合せで指定された一つあるいは複数の無作為抽出カラムのそれぞれのカラムにそれぞれのグループ毎に無作為に値を指定し、前記グループ化処理結果から該指定された値を持つレコードを全て抽出する無作為抽出処理を適用した後、（Ｃ）さらにそれぞれのグループ毎の無作為抽出結果に対して同じく前記（１）で発行された問合せで指定された一つあるいは複数の集計カラムの値に関する集計処理を適用する、からなる処理を含む場合、前記（２-２）の問合せ変換処理は、前記処理（Ａ）、（Ｂ）、（Ｃ）を、それぞれ（Ａ１）処理対象の表に対してまずサンプルグループ化カラムとして問合せで指定されたグループ化カラムを指定し、サンプルグループ化カラムの値に応じて表のレコードをグループに分け、（Ｂ１）前記グループ毎に無作為抽出カラムのそれぞれのカラムに無作為に値を指定し、指定された値を持つレコードを全て抽出する無作為抽出処理を適用した後、（Ｃ１）該無作為抽出処理の結果に対して変換前の問合せにおける分類集計処理を適用する、に変更する処理を備える請求項１記載のデータベース処理システムにおける無作為抽出処理方法。

【請求項５】　前記（２）の問合せ実行管理処理で実行される問合せ処理が、（Ａ）処理対象の二つの表に対して、前記（１）で発行された問合せで指定された一つあるいは複数の結合カラムに等しい値を持つそれぞれの表のレコード同士を結合し、一つのレコードとする結合処理を適用した後、（Ｂ）同じく前記（１）で発行された問合せで指定された一つあるいは複数の無作為抽出カラムのそれぞれのカラムに無作為に値を指定し、指定された値を持つレコードを前記結合結果から全て抽出する無作為抽出処理を適用する、からなる処理を含む場合、前記（２-２）の問合せ変換処理は、前記処理（Ａ）、（Ｂ）を、それぞれ（Ａ２）処理対象のそれぞれの表に対して、無作為抽出カラムをそれぞれの表において共通に含まれるカラムにそれぞれ制限した無作為抽出処理を適用した後、（Ｂ２）それぞれの表からの抽出結果に対して変換前の問合せで指定された結合カラムに関する結合処理を適用する、に変更する処理を備える請求項１記載のデータベース処理システムにおける無作為抽出処理方法。

【請求項６】　前記（２）の問合せ実行管理処理で実行される問合せ処理が、（Ａ）処理対象の表に対して、前記（１）で発行された問合せで指定された一つあるいは複数のサンプルグループ化カラムの値に応じて表のレコードをグループに分けたのち、それぞれのグループ毎に同じく前記（１）で発行された問合せで指定された一つあるいは複数の無作為抽出カラムのそれぞれのカラムに無作為に値を指定し、指定された値を持つレコードを全て抽出する無作為抽出処理を適用した後、（Ｂ）同じく前記（１）で発行された問合せで指定された一つあるいは複数のグループ化カラムの値に応じて表のレコードをグループに分け、前記グループ毎に同じく指定された一つあるいは複数の集計カラムの値に関して集計を行う分類集計処理を適用する、からなる処理を含む場合、前記（２-２）の問合せ変換処理は、前記処理（Ａ）、（Ｂ）を、（Ａ１）処理対象の表に対して無作為抽出カ

ラムからサンプルグループ化カラムを引いた差分を新しい無作為抽出カラムとして用いる無作為抽出処理を適用した後、（Ｂ１）前記（Ｂ）の処理を適用する、に変更する処理を備える請求項１記載のデータベース処理システムにおける無作為抽出処理方法。

【請求項７】　前記（２）の問合せ実行管理処理で実行される問合せ処理が、（Ａ）処理対象の表に対して、前記（１）で発行された問合せで指定された一つあるいは複数の条件評価カラムの値に応じて表のレコードの抽出を行う条件評価処理を適用した後、（Ｂ）同じく前記（１）で発行された問合せで指定された一つあるいは複数の無作為抽出カラムのそれぞれのカラムに無作為に値を指定し、指定された値を持つレコードを前記条件評価処理結果から全て抽出する無作為抽出処理を適用する、からなる処理を含む場合、前記（２-２）の問合せ変換処理は、前記処理（Ａ）、（Ｂ）を、（Ａ１）処理対象の表に対して問合せで指定された一つあるいは複数の無作為抽出カラムを用いた無作為抽出処理を適用した後、（Ｂ１）前記抽出結果に対して同じく前記（１）で発行された問合せで指定された一つあるいは複数の条件評価カラムを用いた条件評価処理を適用する、に変更する処理を備える請求項１記載のデータベース処理システムにおける無作為抽出処理方法。

【請求項８】　前記（２）の問合せ実行管理処理で実行される問合せ処理が、（Ａ）処理対象の表に対して、前記（１）で発行された問合せで指定された一つあるいは複数の射影カラムの抽出を行う射影処理を適用した後、（Ｂ）同じく前記（１）で発行された問合せで指定された一つあるいは複数の無作為抽出カラムのそれぞれのカラムに無作為に値を指定し、指定された値を持つレコードを前記射影処理結果から全て抽出する無作為抽出処理を適用する、からなる処理を含む場合、前記（２-２）の問合せ変換処理が、前記処理（Ａ）、（Ｂ）を、（Ａ１）処理対象の表に対して問合せで指定された無作為抽出カラムを用いた無作為抽出処理を適用した後、（Ｂ１）前記抽出結果に対して問合せで指定された射影カラムを用いた射影処理を適用する、に変更する処理を備える請求項１記載のデータベース処理システムにおける無作為抽出処理方法。

【請求項９】　データベースより所望のデータを抽出するデータベース処理システムにおける無作為抽出処理方法であって、（１）前記データベースに対する問合せを発行する問合せ発行管理処理と、（２）発行された問合せの実行管理を行う問合せ実行管理処理と、（３）前記データベースに対するデータの格納及び格納されたデータの管理を行うデータ管理処理とからなり、前記（１）の問合せ発行管理処理は、（１-１）端末装置からの入力にしたがって問合せ文を生成する問合せ発行処理と、（１-２）問合せ結果および該問合せ結果に対する評価結果を前記端末装置に表示する問合せ結果表示処理とを有し、前記（２）の問合せ実行管理処理は、（２-１）前記（１-１）の問合せ発行処理により発行された問合せ文から無作為抽出処理の挿入された問合せ実行手順および無作為抽出による問合せ結果を評価する問合せ結果評価基準を生成する実行方式生成処理と、（２-２）前記（２-１）の実行方式生成処理により生成された問合せ実行手順を、前記挿入された無作為抽出処理における抽出単位を保存しながら前記問合せをさらに高効率なデータ抽出を行い得る問合せに変換する問合せ変換処理と、（２-３）前記（２-２）の問合せ変換処理で変換した問合せ実行手順にしたがって問合せを実行し、前記（３）のデータ管理処理に対してデータ読み出し要求を発行する問合せ実行処理と、（２-４）前記（２-３）の問合せ実行処理結果を前記（２-１）が生成した問合せ結果評価基準に従って評価し、前記（１）の問合せ発行処理に対して前記問合せ結果および評価結果を受け渡し、該問合せ結果および評価結果に応じて前記（２-３）の問合せ実行処理を制御する問合せ結果評価処理とを有し、前記（３）のデータ管理処理は、（３-１）データベースにデータを格納するデータ格納処理と、（３-２）前記（２-３）の問合せ実行処理が発行したデータ読み出し要求にしたがってデータを読み出すデータ読み出し処理と、を有することを特徴とするデータベース処理システムにおける無作為抽出処理方法。

【請求項１０】　データベースより所望のデータを抽出するデータベース処理システムにおける無作為抽出処理方法であって、（１）前記データベースに対する問合せを発行する問合せ発行管理処理と、（２）発行された問合せの実行管理を行う問合せ実行管理処理と、（３）前記データベースに対するデータの格納及び該データの管理を行うデータ管理処理とを有し、前記（３）のデータ管理処理は、前記（２）の問合せ実行管理処理で実行される問合せ処理に関して、前記（１）で発行された問合せで指定された一つあるいは複数の無作為抽出カラムのそれぞれのカラムに無作為に値を指定し、指定された値を持つレコードを前記問合せ処理結果から全て抽出する無作為抽出処理を含む場合、データとしてのレコードをテーブル化した表からレコードを抽出する際に、無作為抽出カラムに対してハッシュ関数を適用し、あらかじめ無作為決定したハッシュ値を持つレコードを全て抽出した結果を無作為抽出結果とすることを特徴とするデータベース処理システムにおける無作為抽出処理方法。

【請求項１１】　前記（３）のデータ管理処理は、表のレコードを問合せで指定された無作為抽出カラムのハッシュ値の値に応じて互いに排他なレコードの集合であるバケットに分割し、該バケットを外部のレコード記憶装置上の連続領域である１つあるいは複数のブロックに割り付けて格納し、データとしてのレコードをテーブル化した表からレコードを抽出する際に、バケットに含まれる全てのレコードを一つの無作為抽出単位とすること

で、バケットに含まれるレコードの読み出しをレコード記憶装置に対するシーケンシャルアクセスにより実行する請求項１０記載のデータベース処理システムにおける無作為抽出処理方法。

【請求項１２】　データベースより所望のデータを抽出するデータベース処理システムにおける無作為抽出処理方法であって、（１）前記データベースに対する問合せを発行する問合せ発行管理処理と、（２）発行された問合せの実行管理を行う問合せ実行管理処理と、（３）前記データベースに対するデータの格納及び格納されたデータの管理を行うデータ管理処理とを有し、前記（１）の問合せ発行管理処理は、問合せ処理の適切な位置に無作為抽出処理を導入することを指定する問合せ文を発行することを特徴とするデータベース処理システムにおける無作為抽出処理方法。

【請求項１３】　前記（１）の問合せ発行管理処理は、問合せ処理への無作為抽出処理の導入に関する指定に加えて、問合せ処理に要する時間を指定し、前記（２）の問合せ実行管理処理は、指定された問合せ実行時間にしたがって無作為抽出されるデータ量を調節することで問合せで指定された時間内で問合せ処理を終了することを保証する、請求項１２記載のデータベース処理システムにおける無作為抽出処理方法。

【請求項１４】　前記（１）の問合せ発行管理処理は、無作為抽出処理を導入することを指定する問合せ発行処理に加えて、集計処理結果の推定値の精度を指定し、前記（２）の問合せ実行管理処理は、指定された推定値の精度にしたがって無作為抽出されるレコード数を調節することで問合せ発行時に指定された精度を持つ集計処理結果推定値を返す、請求項１２記載のデータベース処理システムにおける無作為抽出処理方法。

【請求項１５】　端末装置からの入力情報に従ってデータベースより所望のデータを抽出するデータベース処理システムであって、（１）前記データベースに対する問合せを発行する問合せ発行管理手段と、（２）発行された問合せの実行管理を行う問合せ実行管理手段と、（３）前記データベースに対するデータの格納及び格納されたデータの管理を行うデータ管理手段とからなり、前記（１）の問合せ発行管理手段は、（１-１）端末装置からの入力にしたがって問合せ文を生成する問合せ発行手段と、（１-２）問合せ結果および該問合せ結果に対する評価結果を端末装置に表示する問合せ結果表示手段とを有し、前記（２）の問合せ実行管理手段は、（２-１）前記（１-１）の問合せ発行手段により発行された問合せ文から無作為抽出処理の挿入された問合せ実行手順および無作為抽出による問合せ結果を評価するための問合せ結果評価基準を生成する実行方式生成手段と、（２-２）前記（２-１）の実行方式生成手段により生成された問合せ実行手順を、前記挿入された無作為抽出処理における抽出単位を保存しながら前記問合せをさらに高効率なデータ抽出を行い得る問合せに変換する問合せ変換手段と、（２-３）前記（２-２）の問合せ変換手段で変換した問合せ実行手順にしたがって問合せを実行し、前記（３）のデータ管理手段に対してデータ読み出し要求を発行する問合せ実行手段と、（２-４）前記（２-３）の問合せ実行処理結果を前記（２-１）で生成した問合せ結果評価基準に従って評価し、前記（１）の問合せ発行手段に対して前記問合せ結果および評価結果を受け渡し、該問合せ結果および評価結果に応じて前記（２-３）の問合せ実行手段を制御する問合せ結果評価処手段と、を有し、前記（３）のデータ管理手段は、（３-１）データベースにデータを格納するデータ格納手段と、（３-２）前記（２-３）の問合せ実行手段により発行されたデータ読み出し要求にしたがってデータベースよりデータを読み出すデータ読み出し手段と、を有することを特徴とするデータベース処理システム。

【請求項１６】　請求項９記載の方法を実行するためのプログラムを記録した記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は関係データベースの問合せ処理方法に係わり、特に大規模なデータベースに対する無作為抽出処理を含む問合せを効率よく実行するための無作為抽出処理方法関する。

【０００２】

【従来の技術】近年、企業内情報処理システムの普及により、取引情報 顧客情報などの様々な業務データがデータベースに蓄積されるようになり、データベースを通じて利用可能となる情報の範囲が急速に拡大しつつある。このときデータベースに蓄積された大規模データを解析し、データの持つ特徴や規則性を抽出することで、ビジネスチャンスの拡大や業務効率向上に役立てることを目的とするデータマイニング処理に対する需要が拡大している。一般にデータマイニング処理では、大規模データが持つ特徴や規則性を様々な観点から解析し抽出するために、データの項目の組み合わせや条件設定を変えて問合せを発行し、繰り返しデータの解析を行う必要がある。しかし、データベースに蓄積されるデータサイズが拡大するにつれて、一回の問合せ処理の所要時間が増大し、効率よくデータの特徴や規則性を抽出することが困難になりつつある。

【０００３】問合せ処理の応答性を向上するための技術としては、文献"ACM SIGMOD International Conference on Management of Data(SIGMOD'96)"(ACM Press発行)のP.205-216に開示されているデータキューブ方式がある。この方式では、問合せを受け付ける前にあらかじめ予想される問合せの処理を行っておき、すでに処理済みの問合せが発行された場合は実際には問合せ処理を行わずに処理済みの結果だけを返すというアプローチである。しかしこのアプローチでは、問合せ結果を事前に用

意しておくために大量の記憶領域を必要とし、事前に処理を行って対処できる問合せの範囲が限られるために、事前に問合せ結果を用意していない問合せに対しては、多大な処理時間が必要とされるという欠点がある。一方、大規模データからの特徴量の算出や規則性の抽出においては、大規模データの持つ傾向や特徴を得ることができればよく、正確な問合せ結果は必要とされないことが多い。そこで問合せ処理時間を大幅に削減する方法として、問合せ処理へ無作為抽出処理を導入し、特徴量や規則性を無作為抽出されたデータから推定することで、処理対象となるデータ量を削減し、応答時間の短縮を図ることが有効である。

【０００４】無作為抽出処理を含む問合せの実行においては、単に無作為抽出によってデータ処理量を減らすだけでなく、問合せの実行前に、問合せの処理の結果を変えずにより実行効率のよい等価な問合せに変換することで大幅に実行時間を短縮する事ができることが重要である。すなわち、無作為抽出処理を問合せ処理のなるべく早い段階で適用して後続の処理の対象となるデータ量を削減することで、問合せ全体の処理量を削減することができる。一般に問合せ処理の対象となるデータの論理構造は図２に示すように表形式（２０）である。この表の横方向をレコード（２１）、縦方向をカラム（２２）という。各レコードの同じカラムは、同じ形式のデータを格納する。レコードの集合である表に対してデータベース処理を適用した後に得られるレコードの集合は、再び表となる。表に対して適用されるデータベース処理とは、表に対する条件評価処理、射影処理、結合処理および分類集計処理を指す。

【０００５】以下ではそれぞれの処理の内容について説明する。まずデータベース処理システムにおける条件評価処理とは、１つ以上の条件評価カラムおよびそれらのカラムの値に設定された条件を指定し、処理対象となる表に含まれるレコードのうち指定されたカラムに関して指定された条件が成り立つようなレコードを抽出し、再び表を構成する処理である。またデータベース処理システムにおける射影処理とは、１つ以上の射影カラムを指定し、処理対象となる表に含まれるそれぞれのレコードについて指定されたカラムだけを抜き出し、再び表を構成する処理である。次にデータベース処理システムにおける結合処理とは、処理対象となる２つの表に共通に含まれる１つ以上の結合カラムを指定し、片方の表に含まれる全てのレコードについて、他方の表に含まれるレコードのうち、結合カラムに同じ値を持つ全てのレコードとの結合を行い、その結果生成される新しいレコードで再び表を構成する処理である。さらにデータベース処理システムにおける分類集計処理とは、１個以上のグループ化カラムおよび１つ以上の集計対象カラムを指定し、処理対象となる表に含まれるレコードを、指定したグループ化カラムの値が同一のレコードを１つのグループとして分類し、それぞれのグループ毎に集計対象カラムの値に関する合計値あるいは平均値などの統計量を計算し、その結果を１つのレコードとして出力する処理である。

【０００６】また本発明では、以下のように定義される無作為抽出処理をデータベースに導入し、無作為抽出処理を含む問合せの変換方法について述べる。本発明のデータベース処理システムにおける無作為抽出処理とは、レコードの集合である表から無作為にレコードを選び出し再び表として構成する処理である。無作為抽出処理において一回の抽出操作で取り出されるレコードの集まりは抽出単位と呼ばれ、一回の抽出処理においては各抽出単位の抽出確率が等しいことを保証される。

【０００７】このようにデータベースに対して発行される問合せは、問合せ対象となる表に対して、上述の様々なデータベース処理を適切な順序で組み合わせて適用することによって構成される。したがって無作為抽出処理を含む問合せの最適化では、無作為抽出処理の無作為抽出性を失わない範囲で問合せを変形し、無作為抽出処理を問合せ処理のなるべく早い段階で適用し後続の処理のデータ量を削減することで、処理時間の短縮を図ることが重要である。従来の無作為抽出処理を含む問合せの変換方式としては文献"International Conference On Very Large Data Bases(VLDB'86)"（Morgan Kaufmann Publishers, Inc. 発行）のP.160-169に開示されている問合せ最適化方式をあげることができる。この方式は、無作為抽出処理と条件評価処理、射影処理、結合処理などの基本的なデータベース処理を含む問合せにおいて、無作為抽出処理の無作為抽出性を保存しつつ処理の適用順序を変更するための問合せ変換処理について開示している。

【０００８】

【発明が解決しようとする課題】従来方式における第一の課題は、問合せ変換処理において、無作為抽出処理の最適化を行おうとする問い合わせに分類集計処理を含む場合に適用することができず、データマイニング応用における問合せの最適化方式としては、限られた有効性しか発揮できないことである。分類集計処理を含む問合せ変換処理において無作為抽出処理の最適化を行う際の問題点は、無作為抽出処理の抽出単位が適切に扱われないことにある。例えば、商品の売上明細情報を顧客ごとに分類し、それぞれの顧客の購入パターンを調べようとした場合、単純に商品の売上明細情報レベルで無作為抽出を行った後、顧客毎ごとの購入パターンの解析を行おうとしても、個々の顧客の商品購入履歴は完全なものを得ることができないために、効率良く購入パターンを解析することができない。これは、顧客毎の購入パターン解析においては顧客毎の購買履歴を抽出単位として抽出し、購入パターン解析を行うべきところを、これを無視した無作為抽出処理を行ったためである。また、第二の

課題は、レコード読み出し処理において、磁気ディスク装置などの記憶装置に格納されたレコードに対して無作為抽出処理を適用する場合、抽出すべきレコードの格納位置がランダムとなるため、記憶装置に対するランダムアクセスが発生し、その結果無作為抽出処理時間が増大することである。さらに、第三の課題は、問合せ発行処理において、問合せ処理時間や問合せ結果の精度などを指定する機構がなかったため、データの一部を無作為抽出により読み出し、ユーザが望むような問合せ結果の推定を行う問合せ発行をユーザが簡便に発行できないことである。

【０００９】本発明の主目的である第１の目的は、上記第一の課題を解決し、無作為抽出処理を含む問合せを効率よく実行するための問合せ実行に際して、抽出単位を考慮した問合せ変換処理としての問合せ最適化を行うことで、分類集計処理を含む問合せに対しても適用可能な無作為抽出処理に関する問合せ変換処理方法を提供することである。本発明の第２の目的は、記憶装置からのレコードの無作為抽出において、記憶装置に対するランダムアクセスを発生させない効率的なレコード格納及び読み出し方法を提供することである。本発明の第３の目的は、問合せの処理実行時間あるいは問合せ結果の精度を指定可能な問合せ発行方法を提供することである。

【００１０】

【課題を解決するための手段】本発明は、上記第１～３の目的を達成するため、それぞれ以下の（１）～（３）の手段を有する。

（１）無作為抽出処理を含む問合せに対して無作為抽出カラムの概念を導入し、抽出単位を考慮した問合せの変換を行うことで分類集計処理を含む問合せをより実行効率の良い問合せに変換する問合せ変換処理を備える。ここで無作為抽出カラムとは、無作為抽出処理における抽出単位を指定するために抽出対象の表に対して指定される一つ以上のカラムである。無作為抽出カラムを利用した無作為抽出処理では、一回の無作為抽出操作において、それぞれの無作為抽出カラムに無作為に値を割り当て、それぞれのカラムに割り当てられた値を持つレコードを表から全て抽出し、無作為抽出結果とする。本発明では、無作為抽出カラムの値が互いに等しいレコードの集まりを一つの抽出単位として扱うように問合せの変換を行う。このことにより例えば先の例では顧客番号を無作為抽出カラムとして指定して同じ顧客番号を持つレコードを単位として抽出を行うことで、完全な顧客購買履歴を得ることができる。ただし以下では、無作為抽出カラムSCが指定されない無作為抽出処理は、レコードのカラム値とは無関係にレコード単位の無作為抽出処理を行い、無作為抽出カラムSCがNULL(空集合)となるような無作為抽出処理では、全件抽出を行うことと定義する。また、本発明における無作為抽出処理では、無作為抽出カラムSCに加えてサンプルグループ化カラムSGCが指定される場合がある。SGCが指定された場合、表のレコードはサンプルグループ化カラムSGCの値に応じてグループに分けられ、それぞれのグループにおいて無作為抽出カラムSCの値が互いに等しいレコードの集まりが一つの抽出単位として抽出されるが、サンプルグループ化カラムが指定された無作為抽出処理では、抽出単位の抽出確率がそれぞれのグループ内で等しいことが保証されると定義する。

【００１１】（２）無作為抽出処理の対象となるレコードのレコード記憶装置に対するレコードの格納及び読み出しにおいて、レコードの無作為抽出カラムにハッシュ関数を適用し、そのハッシュ値に基づいてレコードの格納及び読み出しを行うことでレコード記憶装置に対するランダムアクセスを削減するレコード格納処理及びレコード読み出し処理を備える。

【００１２】（３）問合せ発行時に、問合せ処理所要時間や問合せ結果の推定値の精度を指定し、データベースの規模や問合せの複雑度に応じて無作為抽出抽出されるレコードの量を調節することで、任意の応答時間や精度を持つ問合せ処理を実現する問合せ結果評価処理を備える。

【００１３】本発明による更に他の変形例およびこれを実現するための構成については、実施例において述べる。

【００１４】

【発明の実施の形態】図１に本発明におけるデータベース処理システムの一実施例を示す。まず図１を用いて、本実施例の構成について説明する。本実施例の無作為抽出処理方法は、端末装置１からの入力に従って問合せ文を生成する問合せ発行処理５および問合せ処理結果と評価結果を端末装置に対して表示する問合せ処理結果表示処理６を備える問合せ発行処理２、前記問合せ文から実行可能な中間コードおよび問合せ結果評価基準を生成する実行手順生成処理７および中間コードをより実行効率の良い中間コードに変換する問合せ変換処理８および中間コードにしたがって問合せ処理を行う問合せ処理実行処理９および問合せ結果を前記問合せ評価基準にしたがって評価する問合せ処理結果評価処理１０を備える問合せ実行管理処理３、データとしてのレコードの読み出しを行うレコード読み出し処理１１およびレコードのレコード記憶装置１２への格納を行うレコード格納処理１３を備えるレコード管理処理４、により構成される。これら一連の処理をプログラム化して記録媒体に記録しておけば任意の場所で本発明を利用出来ることになる。

【００１５】以下、図１を用いて本実施例の動作について説明する。まず問合せ発行処理５は、端末装置１からの入力にしたがって問合せ文を生成する。実行手順生成処理７は、前記問合せ発行処理５が生成した問合せ文を参照し、問合せ実行手順と問合せ結果評価基準を生成する。さらに問合せ変換処理８は、前記実行手順生成処理

７が生成した問合せ実行手順をより実行効率の良い問合せ実行手順に変換する。続いて問合せ実行処理９は、前記問合せ変換処理８が変換した問合せ実行手順にしたがって、レコード読み出し処理１１に対してレコード読み出し要求を発行し、読み出したレコードを加工することで問合せ結果を生成する。前記問合せ実行処理９は、問合せ結果評価処理１０から指示されるまで問合せ処理の実行を続ける。レコード読み出し処理１１は、前記問合せ実行処理９からのレコード読み出し、要求に従ってレコード記憶領域１２にレコード格納処理１３によって格納されているレコードを読み出し、読み出されたレコードを前記問合せ実行処理９に対して受け渡す。引き続き問合せ結果評価処理１０は、前記問合せ実行処理９が生成した問合せ処理結果を前記実行手順生成処理７が生成した問合せ結果評価基準に従って評価し、前記問合せ処理結果および前記評価結果を問合せ結果表示処理６に対して送信するとともに、問合せ実行処理を中止すべきかの判断を行い、もし中止すべき場合は前記問合せ実行処理９に対して問合せ処理の中止を指示する。最後に問合せ結果表示処理６は、前記問合せ処理結果評価処理１０が生成した問合せ処理結果とその評価結果を受け取り、これを端末装置１に対して表示する。

【００１６】以下では、前記問合せ発行処理、問合せ手順生成処理、問合せ変換処理、レコード読み出し処理、及び問合せ処理結果評価処理の詳細について具体的な問合せの例を用いて説明する。まず、具体例として図３に示すような３つの表からなるデータベースについて考える。顧客表３１は、顧客番号 顧客分類 名前 住所の４つのカラムからなる。このとき、顧客番号は顧客表のキーカラムであり、表の各レコードごとにユニークな値を持ち、このカラムの値が表におけるレコードを一意に決定する。注文表３２は、注文番号 顧客番号 優先度 注文日の４つのカラムからなる。このとき、注文番号は注文表のキーカラムであり、表の各レコードごとにユニークな値を持ち、このカラムの値が表におけるレコードを一意に決定する。また顧客番号は顧客表のキーカラム顧客番号に対する外部キーであり、顧客表の顧客番号カラムの値の範囲と注文表の顧客番号カラムの値の範囲は一致している。商品表３３は、注文番号 品名 輸送手段 単価の４つのカラムからなる。このとき注文番号は注文表のキーカラム注文番号に対する外部キーであり、注文表の注文番号カラムの値の範囲と商品表の注文番号カラムの値の範囲は一致している。以下では顧客表の顧客番号カラムを例えば顧客表.顧客番号と表わすことにする。

【００１７】本実施例における問合せ発行処理５とは、ＳＱＬ等のデータベース問合せ言語で記述された問合せ文を実行手順生成処理７に受け渡す処理である。例えば前記データベースに対してＳＱＬ形式で記述された以下のような問合せ文を具体例として考える。

```
1:SELECT 顧客区分,優先度,輸送手段,AVG(RANDOM(注文額))
2:FROM SELECT注文番号,顧客区分,優先度,輸送手段,SUM(商品表.単価) AS注文額
3:      FROM顧客表,注文表,商品表
4:      WHERE 顧客表.顧客番号 = 注文表.顧客番号
5:       AND  注文表.注文番号 = 商品表.注文番号
6:      GROUP BY注文表.注文番号,顧客表.顧客区分,注文表.優先度,商品表.輸送手段
7:GROUP BY 顧客区分,優先度,輸送手段;ただし、この問合せ文において文頭の数字は説明のための行番号であり、問合せ文の一部ではない。このとき１行目のキーワードRANDOMは注文額の平均値を算出するのに無作為抽出による推定をすることを指定している。またキーワードRANDOMがキーワードSELECTの直前に指定された場合は、レコードの抽出に無作為抽出を用いることを指定する。
```

【００１８】前記問合せ文の例は、顧客表 注文表 商品表のそれぞれの表から顧客表.顧客番号と注文表.顧客番号の値および注文表.注文番号と商品表.注文番号の値が等しいレコード同士を結合し（3-5行目）、注文表.注文番号,顧客表.顧客区分,注文表.優先度,商品表.輸送手段の４つのカラムの値にしたがってレコードをグループ化し（6行目）、それぞれのグループごとに商品表.単価の合計を求め（２行目）、さらにその結果得られたレコードを顧客区分,優先度,輸送手段の３つのカラムの値にしたがってグループ化し（７行目）、最後にそれぞれのグループごとに注文額の平均値を無作為抽出を用いて推定することを指示する（１行目）。

【００１９】図１２は、図３に示した顧客表 注文表 商品表のそれぞれの表から関連するカラムを抜き出し、顧客表.顧客番号と注文表.顧客番号の値および注文表.注文番号と商品表.注文番号の値が等しいレコード同士を結合した結果を示す。この処理において一つ目の結合処理の結合カラムは{顧客番号}、二つ目の結合処理の結合カラムは{注文番号}である。

【００２０】図１３は、図１２に示した結果を注文表.注文番号,顧客表.顧客区分,注文表.優先度,商品表.輸送手段の４つのカラムの値にしたがってレコードをグループ化し、それぞれのグループごとに商品表.単価の合計を求めた結果を示す。この処理においてグループ化カラムは{注文番号,顧客区分,優先度,輸送手段}、集計対象カラムは{単価}である。

【００２１】図１４は、図１３に示した結果を顧客区分,優先度,輸送手段の３つのカラムの値にしたがってグループ化し（７行目）、それぞれのグループごとに注文額の平均値を算出した結果を示す。この処理においてグループ化カラムは{顧客区分,優先度,輸送手段}、集計対象カラムは{注文額}である。

【００２２】本実施例における実行手順生成処理７とは、問合わせ発行処理５から発行された問合わせ文を、

問合わせ実行処理９において解釈実行が可能となる中間コードに変換する処理である。一般に問合せ文をどのような中間コードに変換するかはデータベース処理システム依存であり、ここではデータベース処理をそれぞれ以下のような中間コードに変換して扱うことにする。すなわち表Ｔに対する条件評価カラムをCCとする条件評価処理をC(CC,T)、表Ｔに対する射影カラムをPCとする射影処理をP(PC,T)、表Sおよび表Ｔに対する結合カラムをJCとする結合処理をJ(JC,S,T)、表Ｔに対する集計カラムをAG、グループ化カラムをGCとする集計処理をA(AG,GC,T)、さらに表Ｔに対する無作為抽出カラムをSC、サンプルグループ化カラムをSGCとする無作為抽出処理をS(SC,SGC,T)と表す。

【００２３】このとき前記問合せ例を中間コードに変換した結果は以下のようになる。

A(注文額,{顧客区分,優先度,輸送手段},
S(無指定,{顧客区分,優先度,輸送手段},
A(商品表.単価,{注文番号,顧客区分,優先度,輸送手段},
J({注文番号},商品表,J({顧客番号},顧客表,注文表))))

このとき本問合せにおいて無作為抽出処理は分類集計処理結果を推定するために導入されているので、無作為抽出処理の無作為抽出カラムSCは無指定,サンプルグループ化カラムSGCは推定しようとする分類集計処理のグループ化カラム{顧客区分,優先度,輸送手段}とする。単純に問合せ結果を無作為抽出抽出する場合、無作為抽出処理の無作為抽出カラムSCは無指定,サンプルグループ化カラムSGCは{}（空集合）とする。このとき前出の問合せ文の例を中間コードに変換した結果を図示すると、図４に示すように木構造となる。問合せ処理の実行では、処理対象となる個々のレコードに対して中間コードの葉に指定された処理から根に指定された処理に向かって順番に処理が適用される。

【００２４】さらに本実施例における実行手順生成処理７では、問合わせ発行処理５から発行された問合わせ文から、問合せ処理結果評価基準を生成し、これを問合せ結果評価処理１０に対して受け渡す。例えば、問合せ発行時に問合せ処理時間が指定された場合は、その指定された処理時間を問合せ結果評価処理に対して受け渡し、あるいはまた集計処理を含む問合せ対して集計結果の精度が指定された場合は、その指定された精度を問合せ結果評価処理に対して受け渡す。問合せ発行時の問合せ処理時間や集計結果の精度の指定方法の一例としては、問合せ文に時間指定あるいは精度指定のためのキーワードを追加することが考えられる。例えば、

SELECT顧客区分, SUM(RANDOM(単価)) AS注文額
FROM顧客表,注文表,商品表
WHERE顧客表.顧客番号＝注文表.顧客番号
AND注文表.注文番号＝商品表.注文番号
WITH IN 2 MINUTES;

なる問合せ文は顧客区分ごとの商品単価合計値を２分以内に求まる範囲で推定することを指定する。また、

SELECT顧客区分, SUM(RANDOM(単価)) AS注文額
FROM顧客表,注文表,商品表
WHERE顧客表.顧客番号＝注文表.顧客番号
AND注文表.注文番号＝商品表.注文番号
WITH 0.99 PRECISION;

なる問合せ文は顧客区分ごとの商品単価合計値を９９％の精度で推定することを指定する。このとき上記２つの問合せ文において、時間指定あるいは精度指定のキーワードのみが指定された場合、問合せ発行処理において無作為抽出処理の挿入位置を自動決定し、RANDOMキーワードを補完することで端末装置からの問合せ文発行の手間を軽減することも可能である。

【００２５】本実施例における問合せ変換処理８では、前記実行手順生成処理７によって生成された前記中間コードに導入されている無作為抽出処理と、その直前に適用される問合せ処理との間で処理の適用順序の交換を行い、より実行効率の良い中間コードに変換する。このとき前記問合せ変換処理では、無作為抽出処理を含む問合せに対して、無作為抽出カラムを用いて抽出単位を保存するような問合せの変形を行うことで、無作為抽出処理の無作為性を保存しつつ、より実行効率のよい問合せに変換する。

【００２６】まず前記問合せ例の中間コードに対して問合せ変換を適用した場合の様子を以下に示す。図５に示すように、中間コードに挿入されている無作為抽出処理の無作為抽出カラムの値を、直前の分類集計処理-1のグループ化カラムの値{注文番号,顧客区分,優先度,輸送手段}に変更して無作為抽出処理と分類集計処理の適用順序を変更する。この変更により図１５に示すように分類集計処理-1適用前の表から各無作為抽出カラムに無作為に値が割り当てられ、例えば{注文番号：注文１,顧客区分：建築,優先度：高,輸送手段：トラック}を満たすレコードが全て抽出されるようになる。このとき、分類集計処理後の表においてグループ化カラムはユニークカラムであるので、グループ化カラムの値と個々のレコードは１対１に対応する。したがって、分類集計処理後の表に対してレコード単位で無作為抽出を行うことと分類集計処理前にグループ化カラムの値を無作為に指定して無作為抽出を行うこととは等価である。したがって上記のような処理の適用順序の交換によっても無作為抽出処理の無作為性は失われない。

【００２７】次に、図６に示すように、無作為抽出処理の無作為抽出カラムを変換前の無作為抽出カラムからサンプルグループ化カラムを引いた差分、すなわち{注文番号}に変更して無作為抽出処理を適用するようにする。この変更により図１６に示すように分類集計処理-1適用前の表から、無作為抽出カラムに無作為に値が割り当てられ、例えば{注文番号：注文1}を持つレコードが全て抽出されるようになる。このとき、無作為抽出前の

表をサンプルグループ化カラム{顧客区分,優先度,輸送手段}の値に応じてグループ分けを行うと、それぞれのグループではサンプルグループ化カラムの値は互いに等しく、無作為抽出カラム{注文番号,顧客区分,優先度,輸送手段}にそれぞれ値を割り当てレコードの抽出を行っても、このうちレコードの指定に有効なのは{注文番号}のみである。したがって、上記のような処理の適用順序の交換によって無作為抽出処理の無作為性が失われることはない。

【００２８】さらに、次の結合処理-2と無作為抽出処理の適用順序を変更するために、図７に示すように結合前のそれぞれの表に対して無作為抽出処理を分配する。この変更により集計処理適用前の表から、無作為抽出カラムに割り当てられた値{注文番号：注文1}を持つレコードが全て抽出されるようになる。このとき、結合処理後の表から無作為抽出カラムに指定した値を持つレコードを抽出して得られるレコードの集合と、件都合処理雨のの表から無作為抽出カラムに指定した値を持つレコードを抽出して得られるレコードの集合は１対１に対応するので、処理の適用順序の交換によって無作為抽出処理の無作為性は失われない。

【００２９】さらに続く結合処理-1においても、図８に示すように結合前のそれぞれの表に対して無作為抽出処理を分配し、適用順序を交換する。この変更により集計処理適用前の表から、無作為抽出カラムに割り当てられた値{注文番号：注文1}を持つレコードが全て抽出されるようになる。この適用順序の交換も前述の理由と同じ理由で可能である。ただし顧客表はカラムとして{注文番号}を含まないので、顧客表に分配される無作為抽出処理の無作為抽出カラムはNULLであり、顧客表に対しては全件抽出を行う。以上の問合せ変換により、前記問合せ例の中間コードは図９のように変換される。

【００３０】以下では、各種問合せ処理と無作為抽出処理との間の処理順序の交換方法についてまとめて述べる。無作為抽出処理と条件評価処理との交換は、単純に両者の適用順序を入れ替えればよい。すなわちS(SC,SGC,C(CC,T))≡C(CC,S(SC,SGC,T))が成り立つ。このとき≡は、両辺の操作が無作為抽出処理として等価であることを示す。両辺の操作が無作為抽出処理として等価であるためには、

(i)抽出単位が両辺の処理で一致している。

(ii)両辺の処理において各抽出単位の抽出確率が保存されているという２点あるいは(i)抽出単位が両辺の処理で一致している(ii)変換後の問合せにおいて各無作為抽出処理の抽出単位が互いに独立で等しい抽出確率で抽出されるという２点を示せば良い。したがって上記の両辺の処理が等価であることは、以下の２点からわかる。

(i)一回の抽出処理において抽出されるレコードは、レコードが分類されたサンプルグループのサンプルグループ化カラムSGCの値をsgc、無作為抽出カラムSCに指定された値をscとして、表Tに含まれるレコードのうち、SC=scおよびSGC=sgcおよび条件評価カラムCCに指定された条件を満たすレコードであり、両辺の抽出単位は等しい。

(ii) 変換前後の問合せにおいてSCおよびSGCの値の組と問合せ変換前後での抽出単位は１対１対応しており、SGCに対してSCの値が無作為決定されるならば、問合せ変換において各抽出単位の抽出確率は保存されている。

【００３１】SCが無指定の場合、レコード単位で無作為抽出が行われることになるが、条件評価処理によって処理前のレコードと処理後のレコードは１対１に対応しており、変換前後の問合せにおいて抽出単位は等しく、表Tのレコードのうち条件を満たすレコードに関して問合せ変換の前後で抽出確率は保存されている。SC=NULLの場合、無作為抽出処理は全件抽出なので両辺の操作は一致する。無作為抽出処理と射影処理との交換は、単純に両者の適用順序を入れ替えればよい。すなわち、S(SC,SGC,P(PC,T))≡P(PC,S(SC,SGC,T))が成り立つ。ただし左辺の無作為抽出操作が可能であるためには、SC⊆PCが必要である。

【００３２】このとき上記の両辺の処理が等価であることは、以下の２点からわかる。

(i)一回の抽出処理において抽出されるレコードは、レコードが分類されたサンプルグループのサンプルグループ化カラムSGCの値をsgc、無作為抽出カラムSCに指定された値をscとして表Tに含まれるレコードのうち、SC=scおよびSGC=sgcを満たすレコードであり、両辺の抽出単位は等しい。

(ii)SCおよびSGCの値の組と抽出単位は１対１に対応しており、SGCに対してSCの値が無作為決定されるならば、問合せ変換において各抽出単位の抽出確率は保存されている。

【００３３】SCが無指定の場合、射影処理によって処理前のレコードと処理後のレコードは１対１に対応しており、変換前後の問合せにおいて抽出単位は等しく、レコード毎の抽出確率も保存されている。SC=NULLの場合、無作為抽出処理は全件抽出なので両辺の操作は一致する。無作為抽出処理と分類集計処理との交換は、単純に両者の適用順序を入れ替えればよい。すなわち、S(SC,SGC,A(AC,GC,T))≡A(AC,GC,S(SC,SGC,T))が成り立つ。このとき上記の両辺の処理が等価であることは、以下の２点からわかる。

(i)一回の抽出処理において抽出されるレコードは、レコードが分類されたサンプルグループのサンプルグループ化カラムSGCの値をsgc、無作為抽出カラムSCに指定された値をsc、分類集計処理において分類されたグループのグループ化カラムGCの値をgcとして表Tに含まれるレコードのうちSC=scおよびSGC=sgcおよびGC=gcを満たすレコードであり、両辺の抽出単位は等しい。

(ii)SCおよびSGCおよびGCの値の組と抽出単位は１対１

に対応しており、 SGCに対してSCの値が無作為決定されるならば、問合せ変換において各抽出単位の抽出確率は保存されている。

【００３４】SCが無指定の場合、SCに分類集計処理のグループ化カラムを指定し、問合せの変換を行う。グループ化処理がない場合は、分類集計処理の結果はレコード一つであるので、無作為抽出処理は行わず、全件抽出を行う。SC=NULLの場合、無作為抽出処理は全件抽出なので両辺の操作は一致する。無作為抽出処理と結合処理との交換は、無作為抽出カラムＳＣ及びサンプルグループ化カラムＳＧＣをそれぞれの表に含まれるカラムに制限し、両者の適用順序を交換すればよい。すなわちS(SC,SGC,J(JC,S,T))≡J(JC,S(SC/S,SGC/S,S),S(SC/T,SGC/T,T))が成り立つ。

【００３５】両辺の処理が等価であることは、以下の２点からわかる。

(i)一回の抽出処理において抽出されるレコードは、レコードが分類されたサンプルグループのサンプルグループ化カラムSGCの値をsgc、無作為抽出カラムSCに指定された値をscとして、右の処理で抽出されるレコードは、表ＳでSC/S=sc/SおよびSGC/S=sgc/S,表ＴでSC/T=sc/TおよびSGC/T=sgc/Tを満たすレコードでJC同士の等しいもので、左の処理でも抽出される。右の処理で抽出されないレコードは、上記のいずれかの条件を満たさないので、左の処理でも抽出されない。よって両辺の抽出単位は等しい。

(ii)SCおよびSGCの値の組と抽出単位は１対１対応しており、 SGCに対してSCの値が無作為決定されるならば、問合せ変換において各抽出単位の抽出確率は保存されている。

【００３６】SCの指定がない場合、J(JC,S,T)のキーカラムをKCとして、S(指定無し,SGC,J(JC,S,T))≡ S(KC,SGC,J(JC,S,T))とする。J(JC,S,T)にキーカラムが存在しない場合は、ジョインの属性値に応じた重み付けをして分配。すなわち、表Ｔにおける属性Ｘの属性値xiの出現比率を|T.xi|として、表Ｓからの属性値xiを持つレコードの抽出確率を|T.xi|/|T.x|maxとして無作為抽出を行う。S(指定無し,SGC,J(JC,S,T))≡S(指定無し,SGC,J(JC,Select(|T.xi|/|T.x|max,S(指定無し,S)),T))なる変換で、抽出単位は変わらない。また、各抽出単位の抽出単位の抽出操作は独立であり、その抽出確率も

1/|S|*|T.xi|/|T.x|max *1/|T.xi|=1/|S||T.x|max

で互いに等しい。SC=NULLの場合、無作為抽出処理は全件抽出なので両辺の操作は一致する。また、無作為抽出処理に関してサンプルグループ化カラムSGCが無作為抽出カラムSCに含まれる場合(SC⊇SGC)、無作為抽出カラムを変換前の無作為抽出カラムからサンプルグループ化カラムを引いた差分に置き換えることができる。例えば、図１７に示すように、SC={X,Y},SGC={Y}とすると、この無作為抽出処理における抽出単位はカラムＸ,Ｙの値の等しいレコードとなるが、表のレコードをカラムＹの値に関してグループ分けした後、カラムＸ,Ｙの値に関して無作為抽出を行った場合と、まず表のレコードを無作為抽出カラムSCから無作為抽出カラムSGCを引いたカラムＸの値に関して無作為抽出を行った後、カラムＹの値に関してグループ化を行った場合も、それぞれのグループにおいて抽出単位と抽出確率が等しいことに変化は生じない。すなわちS(SC,SGC,T)≡S(SC-SGC,SGC,T)が成り立つ。

【００３７】両辺の処理が等価であることは、以下の２点からわかる。

(i)左辺の処理において、一回の抽出処理において抽出されるレコードの一つをｒとして、ｒの無作為抽出カラムSCの値をsc、サンプルグループ化カラムSGCをsgcとすると、SGC=sgcおよびSC-SGC=sc-sgcが成り立つので、SCおよびSGCの同じ値の割り当てによってレコードｒは右辺の処理においても抽出される。また逆に、左辺の処理において、一回の抽出処理において抽出されるレコードの一つをｒとして、ｒの無作為抽出カラムSC-SGCの値をsc-sgc、サンプルグループ化カラムSGCをsgcとすると、SGC=sgcおよびSC=scが成り立つので、SCおよびSGCの同じ値の割り当てによってレコードｒは左辺の処理においても抽出される。以上より、 SCおよびSGCの同じ値の割り当てに対して同じレコードが抽出されることからその抽出単位は等しい。

(ii) SCおよびSGCの値の組と抽出単位は１対１対応しており、SGCに対してSCの値が無作為決定されるならば、問合せ変換において各抽出単位の抽出確率は保存されている。

【００３８】本実施例におけるレコード読み出し処理１１は、前記問合せ実行処理９が発行するレコード読み出し要求にしたがって、レコード記憶装置１２に格納されたレコードの読み出しを行う。図１１に本実施例におけるレコード格納の様子を示す。本実施例におけるレコード記憶装置１２へのレコードの格納では、レコード格納処理１３を用いて、レコードの１個以上のレコードのカラムをあらかじめ分割カラムBCとして指定し、指定された分割カラムに対してハッシュ関数１１１を適用し、その値に応じてレコードをバケット１１２と呼ばれるグループに分類して格納しておく。ハッシュ分割に用いるハッシュ関数１１１としては、文献“PRINCIPLES OF DATABASE AND KNOWLEDGE-BASE SYSTEMS”,J.D.Ullman著 Computer Science Press発行 P.358-360に開示されている分割ハッシュ関数等を用いることで、分割カラム毎にハッシュ値を指定したレコード読み出しが行えるようになる。ただし、分割カラムの指定がない場合は、全てのレコードを一つのバケットに格納する。それぞれのバケットに対しては、レコード記憶装置内の連続領域であるブロック１１３が必要に応じて割り当てられ、レコードはそれらのブロックに格納される。したがって同じハッシ

ュ値を持つレコードの読み出しにおいては、レコード記憶装置に対するランダムアクセスは発生しない。

【００３９】本実施例におけるレコード記憶装置１２からのレコードの読み出しでは、レコード読み出し処理直後のデータベース処理の内容によって以下の４つの方式を使い分ける。

方式１：直後のデータベース処理が無作為抽出処理以外あるいは直後の無作為抽出処理の無作為抽出カラムがNULLの場合、通常の読み出し処理を行う。すなわち、全てのバケットに含まれるレコードを全て読み出す。

方式２：直後の無作為抽出処理の無作為抽出カラムが無指定の場合、レコード単位の無作為抽出処理を行う。すなわち、一回の読み出し要求に対してレコード格納領域１２に格納されるレコードを無作為に１つ決定し、レコード読み出しを行う。

方式３：直後の無作為抽出処理の無作為抽出カラムSCとレコードの分割カラムBCとの間に共通部分がない場合、ハッシュ関数を適用したレコード読み出し処理を行う。すなわち、図１０に示すようにレコード読み出し時に無作為抽出カラムのそれぞれのカラムについてハッシュ値を無作為に指定し、レコード記憶装置に格納されたレコードの無作為抽出カラムSCに対してハッシュ関数１０１を適用し、指定されたハッシュ値を持つレコードだけを抽出する。分割カラムの指定が無い場合も、このレコード読み出し方式を利用する。

方式４：直後の無作為抽出処理の無作為抽出カラムとレコードの分割カラムとの間に共通部分がある場合、バケット分割されたレコードを利用したレコード読み出し処理を行う。すなわち、図１１に示すようにレコード読み出し時に無作為抽出カラムのそれぞれのカラムについてハッシュ値を無作為に指定し、レコード記憶装置１２に格納されたレコードの無作為抽出カラムSCと格納分割カラムBCの共通部分SC∩BCに関しては、指定されたハッシュ値を持つバケット１１２のブロック１１３に格納されたレコードを読み出し、さらに読み出されたレコードの無作為抽出カラムSCと格納分割カラムBCの差分SC-BCに対してハッシュ関数１１４を適用し、指定されたハッシュ値を持つレコードだけを抽出する。

【００４０】本実施例におけるレコード記憶装置１２は、表形式にまとめられたレコードを格納する。レコードをバケットに分割して格納する場合、レコードの分割カラムのハッシュ値によってそれぞれのバケットに割り当てられたレコードをレコード記憶装置上の連続領域であるブロックに割り当てて格納することで、レコードアクセスがシーケンシャルアクセスとなり、バケット読み出し効率を向上させることができる。

【００４１】本実施例における問合せ結果評価処理１０では、前記実行手順生成処理７が生成した問合せ評価基準に基づき、前記問合せ実行処理９での問合せ結果を評価し、問合せ処理の実行制御を行う。前記実行手順生成処理７が生成した問合せ評価基準において時間を指定された場合、問合せ処理評価処理では、問合せ実行処理から問合せ結果を受け取る毎に問合せ実行時間と指定された時間を比較し、問合せ実行時間が指定された時間を超過していた場合は、問合せ実行処理に対して問合せ処理の中止を指示する。前記実行手順生成処理７が生成した問合せ評価基準において精度を指定された場合、問合せ処理評価処理では、問合せ実行処理から問合せ結果を受け取る毎に、問合せ結果の推定値の精度を計算し、推定値の精度が指定された精度を超過していた場合は、問合せ実行処理に対して問合せ処理の中止を指示し、そのときの推定値と精度を返す。

【００４２】このとき推定値の精度は、文献「統計学辞典」竹内啓著　東洋経済新報社出版P.243-247,252-254に開示されている、無作為抽出法および集落抽出法に関する推定方法によって算出することができる。また上記実施例において、問合せ発行処理あるいは問合せ変換処理のみをそれぞれ単独に備える問合せ処理方式を用いてデータベース処理システムを構成することも可能である。

【００４３】

【発明の効果】本発明による問合せ変換処理では、無作為抽出処理と他の問合せ処理との間の適用順序の交換において、無作為抽出処理の抽出単位を考慮した問合せの変換を行うことにより、抽出単位を考慮しない従来の問合せ変換処理と比べ、集計処理を含む問合せに対しても適用することができ、さらに広い範囲の問合せの効率向上を図ることができる。また問合せ発行時に、問合せ処理所要時間や問合せ結果の推定値の精度を指定し、データベースの規模や問合せの複雑度に応じて無作為抽出されるレコードの量を調節することで、任意の応答時間や精度を備えた問合せ結果を簡便に得ることができる。さらに、表からバケット単位で無作為抽出を行い集計処理を行う際に、集計対象となるバケット当りのレコード数が十分多い場合に、集計結果の精度計算を単純な無作為抽出処理の場合の精度計算式で近似することで、レコード分割に伴う新たな統計量を必要とせず集計結果の精度計算を単純な無作為抽出処理の場合の精度計算式で近似することができる。

【図面の簡単な説明】

【図１】本発明によるデータベース処理システムの実施例の概要図である。

【図２】表の構成に関する説明図である。

【図３】データベースの例の説明図である。

【図４】問合せ例の変換前の中間コードの説明図である。

【図５】問合せ例での無作為抽出処理と分類集計処理-１の適用順序交換の説明図である。

【図６】問合せ例での無作為抽出処理の無作為抽出カラム変換の説明図である。

【図７】問合せ例での無作為抽出処理と結合処理-２の適用順序交換の説明図である。
【図８】問合せ例での無作為抽出処理と結合処理-１の適用順序交換の説明図である。
【図９】問合せ例の変換後の中間コードの説明図である。
【図１０】レコード記憶装置からのレコード読み出し処理の説明図である。
【図１１】レコード記憶装置へのレコード格納処理の説明図である。
【図１２】問合せ例での結合処理結果の例の説明図である。
【図１３】問合せ例での集計処理-1の処理結果の説明図である。
【図１４】問合せ例での集計処理-2の処理結果の説明図である。
【図１５】無作為抽出処理と分類集計処理-１との問合せ変換により抽出されるレコードの説明図である。
【図１６】無作為抽出処理の無作為抽出カラムの変換により抽出されるレコードの説明図である。
【図１７】無作為抽出処理の無作為抽出カラムの変換方法の説明図である。

【符号の説明】
１　端末装置
２　問合せ発行処理
３　問合せ実行管理処理
４　レコード管理処理
５　問合せ発行処理
６　問合せ結果表示処理
７　実行手順生成処理
８　問合せ変換処理
９　問合せ実行処理
１０　問合せ結果評価処理
１１　レコード読み出し処理
１２　レコード記憶装置
１３　レコード格納処理
２０　表
２１　レコード
２２　カラム
３１　顧客表
３２　注文表
３３　商品表

【図２】

図2

20 表　　カラム 21

注文番号	地域	数量	金額
０１	東京	１０	２０００
０２	東京	２０	１０００
０３	東京	３０	３０００
０４	大阪	１０	５０００
０５	東京	４０	３０００
０６	名古屋	２０	２０００
０７	東京	１０	４０００
０８	大阪	３０	１０００
０９	名古屋	２０	４０００
１０	東京	４０	３００

22 レコード

【図３】

図3

顧客表 31

顧客番号	顧客分類	名前	住所
顧客１	建設	青山	東京
顧客２	食品	井上	大阪
顧客３	建設	上田	名古屋

注文表 32

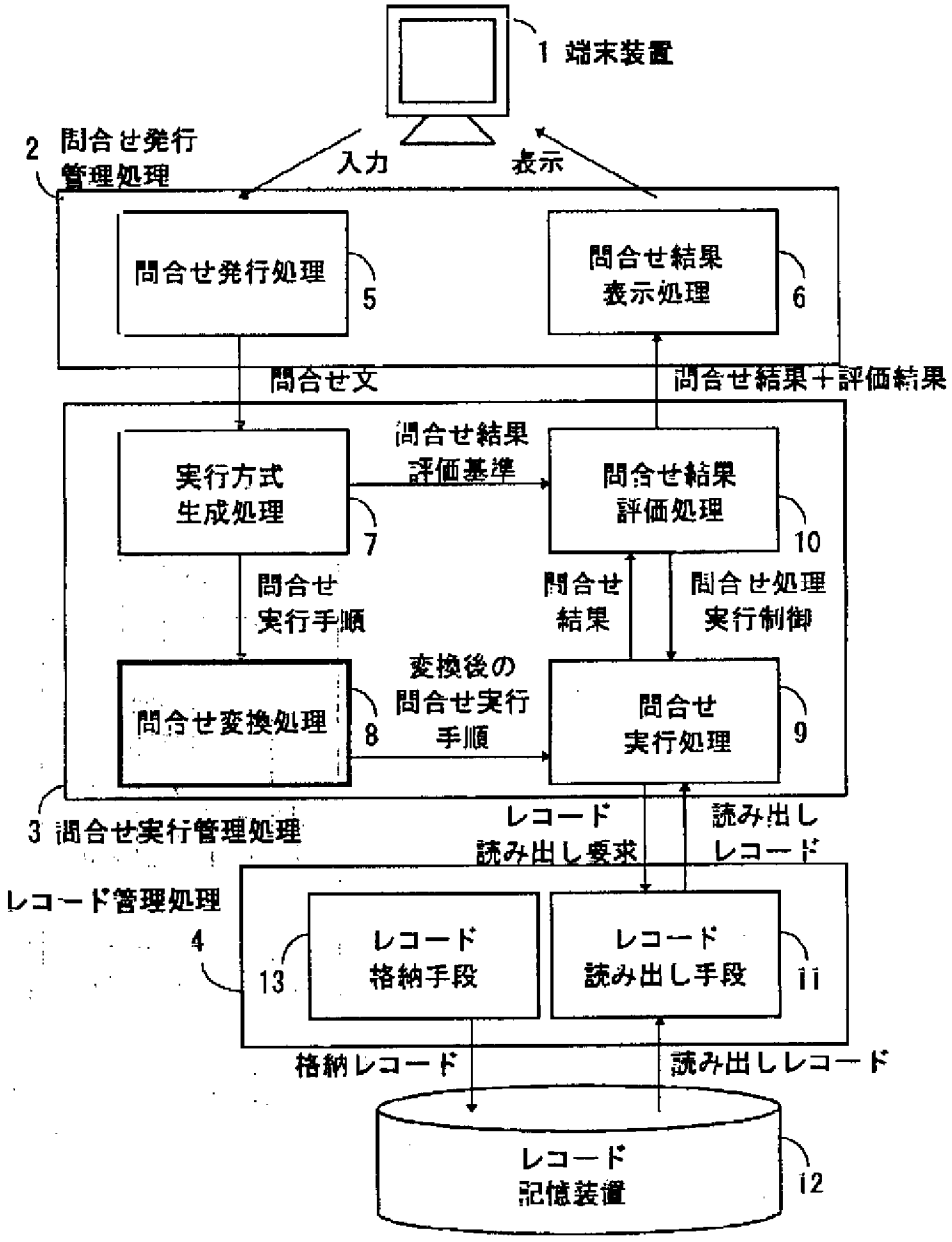
注文番号	顧客番号	優先度	注文日
注文１	顧客１	高	5/14
注文２	顧客２	低	5/16
注文３	顧客３	高	5/19
注文４	顧客１	低	5/20

商品表 33

注文番号	品名	輸送手段	単価
注文１	セメント	貨車	3500
注文１	鉄板	トラック	6000
注文１	木材	トラック	7000
注文２	小麦	船便	500
注文２	牛肉	船便	800
注文３	セメント	貨車	7500
注文４	木材	船便	2000

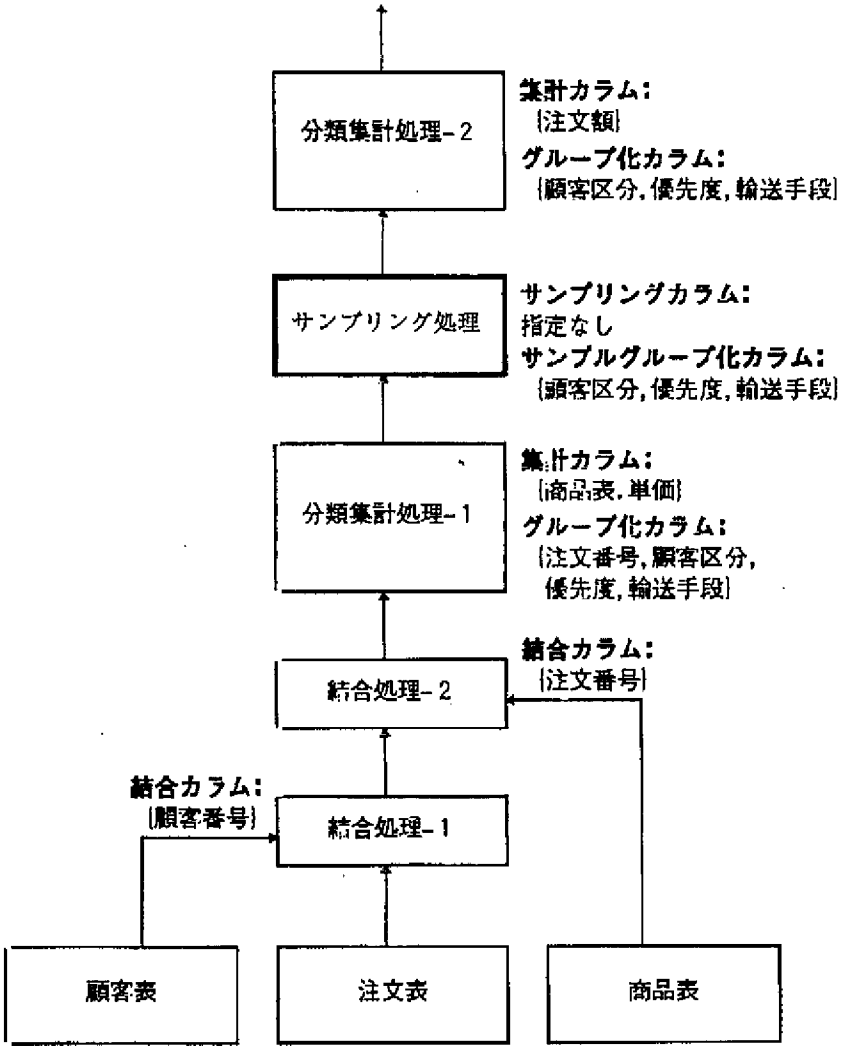
【図1】

図1


1 端末装置
2 問合せ発行管理処理
入力
表示
問合せ発行処理
5
問合せ結果表示処理
6
問合せ文
問合せ結果＋評価結果
実行方式生成処理
7
問合せ結果評価基準
問合せ結果評価処理
10
問合せ実行手順
問合せ結果
問合せ処理実行制御
問合せ変換処理
8
変換後の問合せ実行手順
問合せ実行処理
9
3 問合せ実行管理処理
レコード読み出し要求
読み出しレコード
レコード管理処理
4
13
レコード格納手段
レコード読み出し手段
11
格納レコード
読み出しレコード
レコード記憶装置
12

【図４】

図４



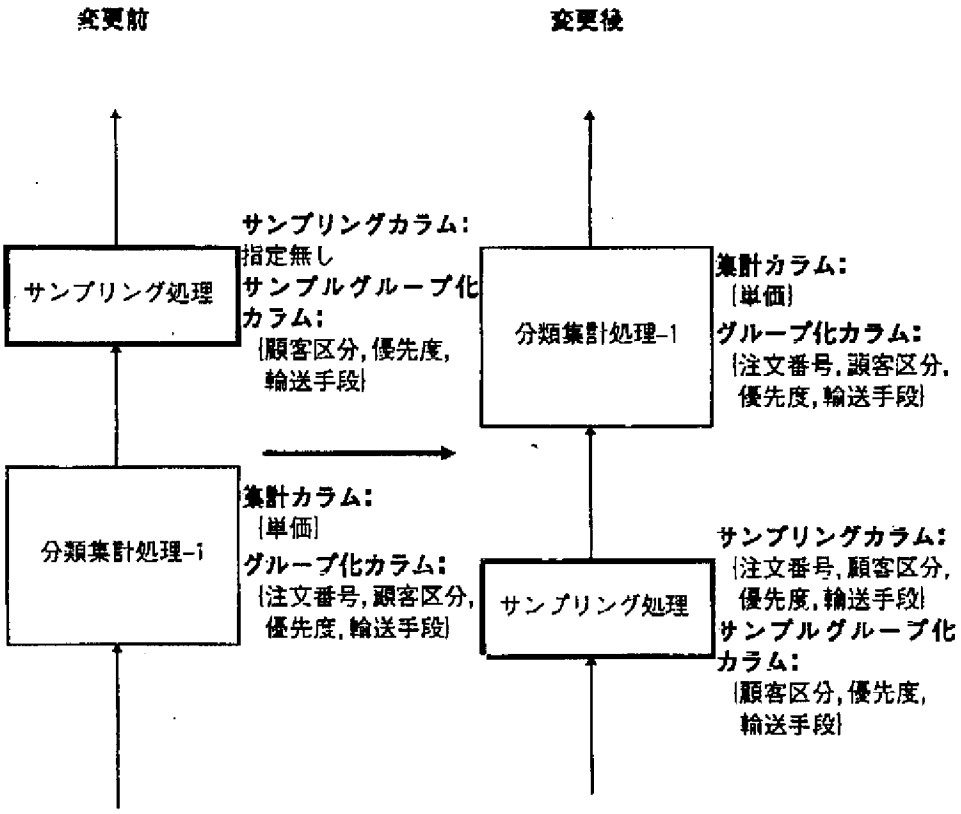
【図１４】

図１４

顧客区分	優先度	輸送手段	平均注文額
商数	高	貨車	5500
建数	高	トラック	8000
食品	低	船便	1300
建設	低	船便	2000

【図5】

図5



【図12】

図12

注文番号	顧客区分	優先度	輸送手段	単価
注文1	建設	高	貨車	3600
注文1	建設	高	トラック	6000
注文1	建設	高	トラック	2000
注文2	食品	低	船便	500
注文2	食品	低	船便	800
注文3	建設	高	貨車	7500
注文4	建設	低	船便	2000

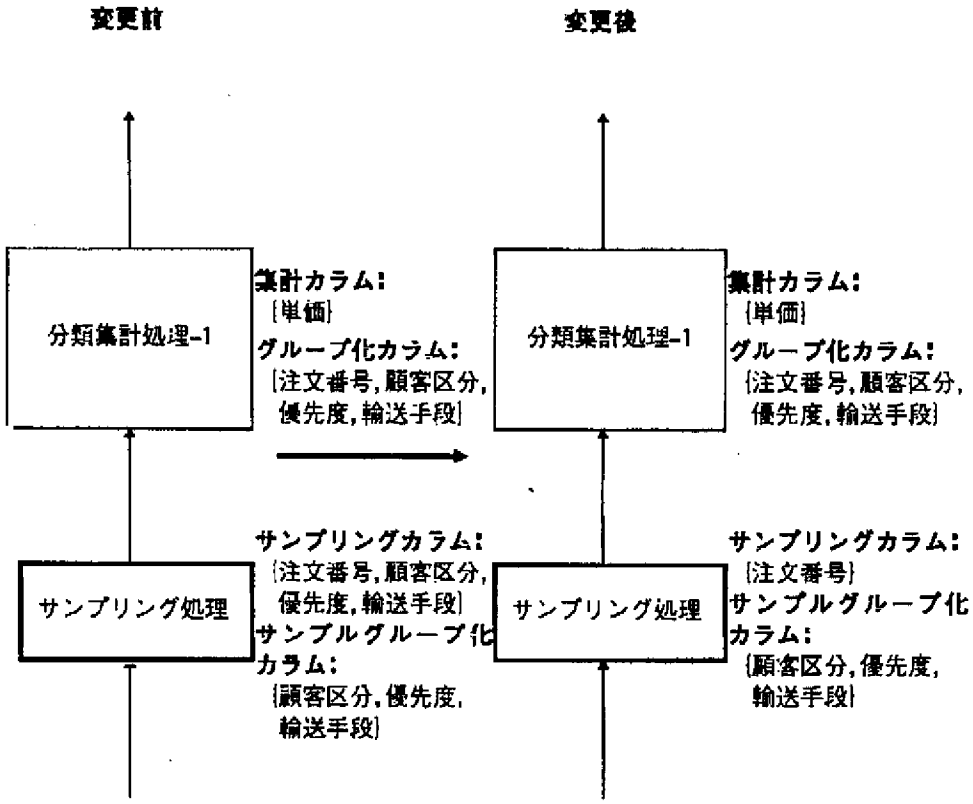
【図13】

図13

注文番号	顧客区分	優先度	輸送手段	注文額
注文1	建設	高	貨車	3600
注文1	建設	高	トラック	8000
注文2	食品	低	船便	1300
注文3	建設	高	貨車	7500
注文4	建設	低	船便	2000

【図6】

図6



【図15】

図15

注文番号	顧客区分	優先度	輸送手段	注文額
注文1	建設	高	貨車	3500
注文1	建設	高	トラック	6000
注文2	食品	低	船便	1300
注文3	建設	高	貨車	7500
注文4	建設	低	船便	2000

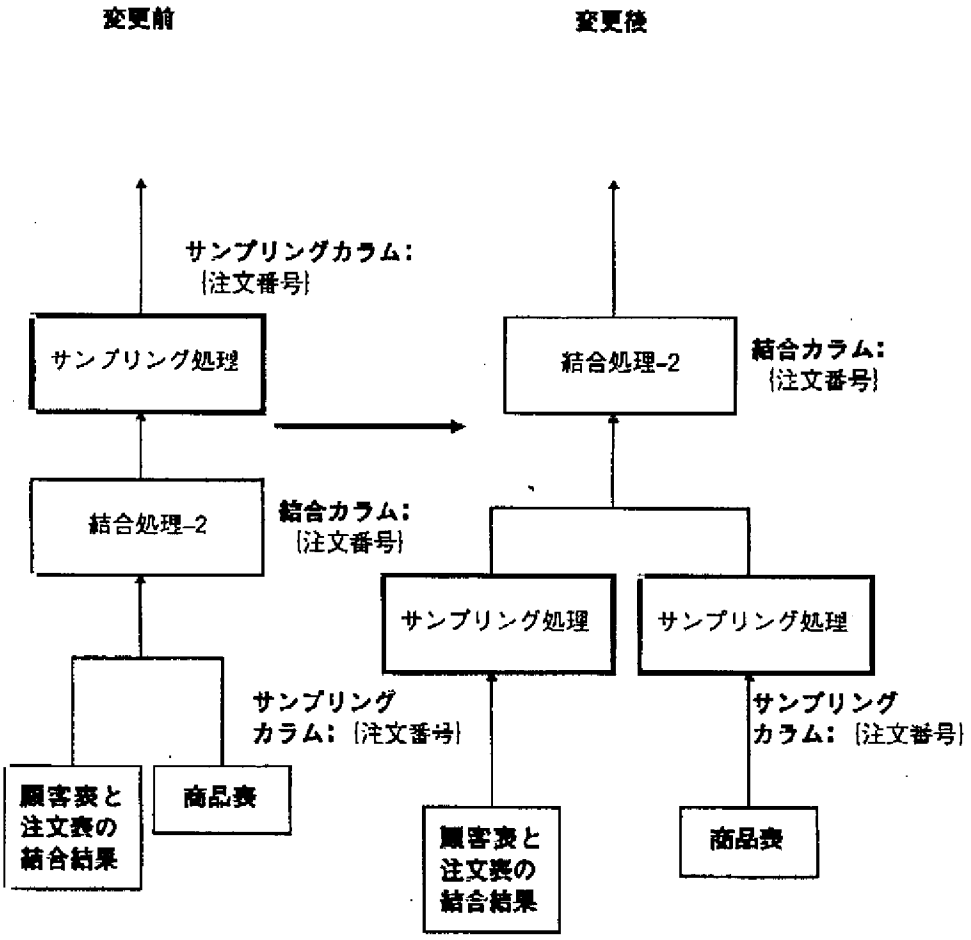
【図16】

図16

注文番号	顧客区分	優先度	輸送手段	単価
注文1	建設	高	貨車	3500
注文1	建設	高	トラック	6000
注文1	建設	高	トラック	2000
注文2	食品	低	船便	500
注文2	食品	低	船便	800
注文3	建設	高	貨車	7500
注文4	建設	低	船便	2000

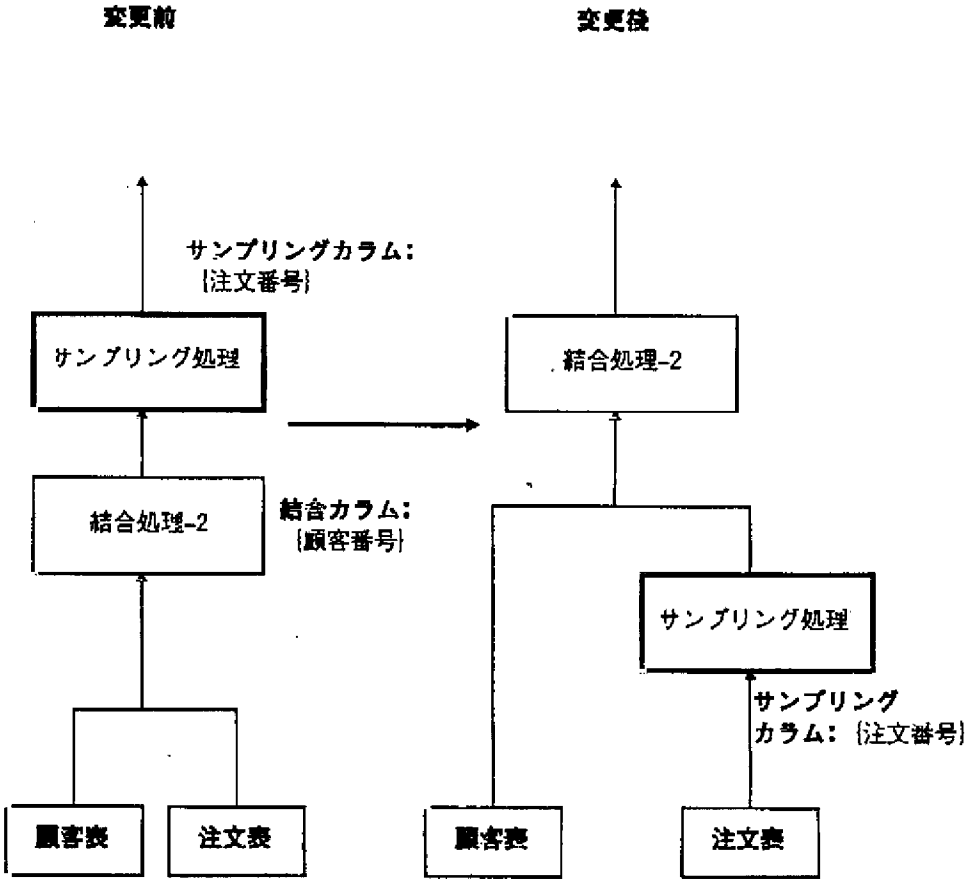
【図7】

図 7


変更前
変更後
サンプリングカラム:
{注文番号}
サンプリング処理
結合処理-2
結合カラム:
{注文番号}
結合処理-2
結合カラム:
{注文番号}
サンプリング処理
サンプリング処理
サンプリング
カラム: {注文番号}
サンプリング
カラム: {注文番号}
顧客表と
注文表の
結合結果
商品表
顧客表と
注文表の
結合結果
商品表

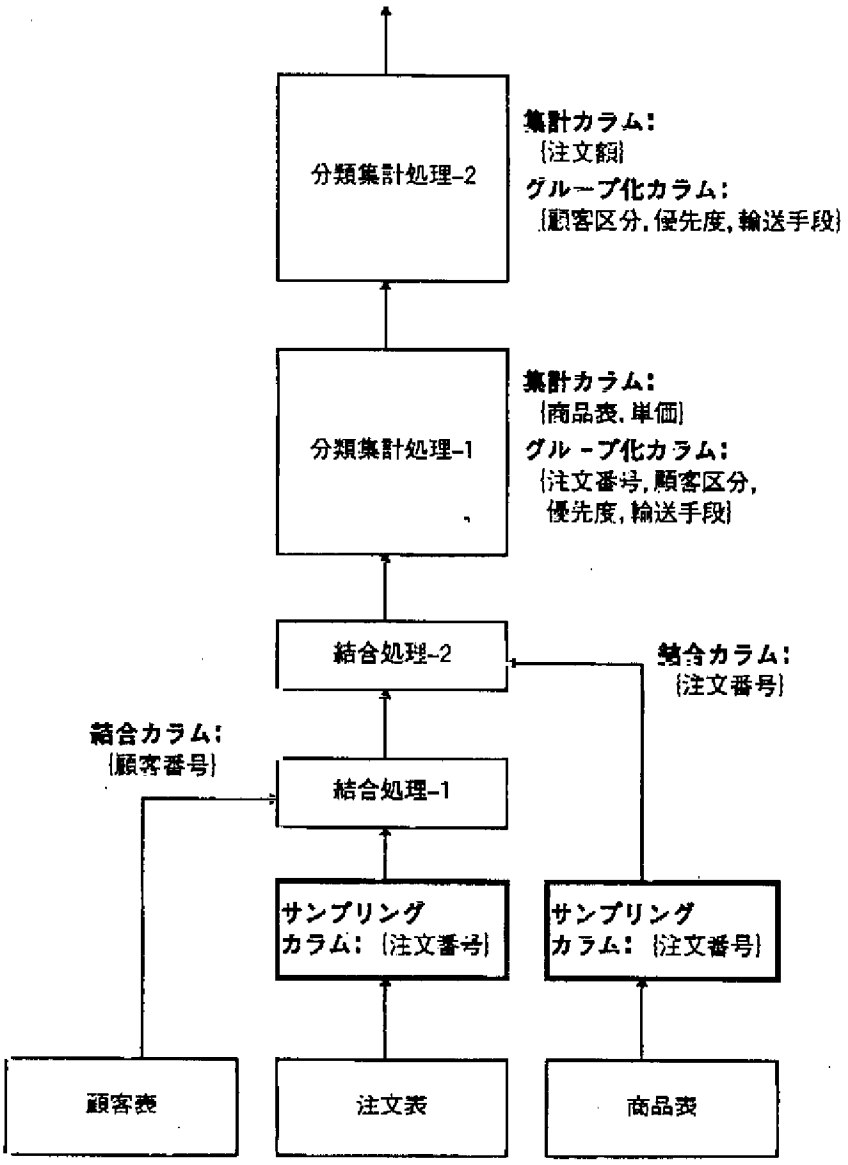
【図8】

図8


変更前
変更後
サンプリングカラム：
{注文番号}
サンプリング処理
結合処理-2
結合処理-2
結合カラム：
{顧客番号}
サンプリング処理
サンプリング
カラム：{注文番号}
顧客表
注文表
顧客表
注文表

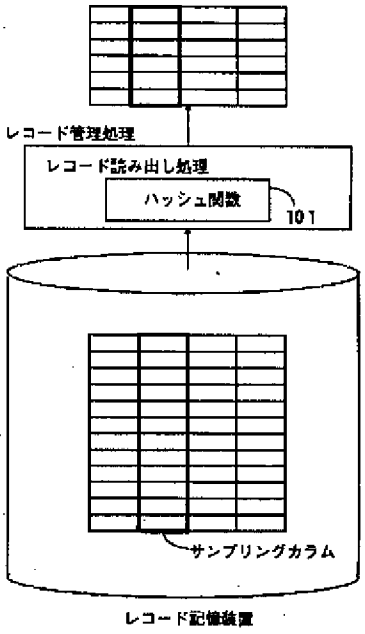
【図9】

図9


分類集計処理-2
集計カラム:
{注文額}
グループ化カラム:
{顧客区分,優先度,輸送手段}
分類集計処理-1
集計カラム:
{商品表.単価}
グループ化カラム:
{注文番号,顧客区分,
優先度,輸送手段}
結合処理-2
結合カラム:
{注文番号}
結合カラム:
{顧客番号}
結合処理-1
サンプリング
カラム:{注文番号}
サンプリング
カラム:{注文番号}
顧客表
注文表
商品表

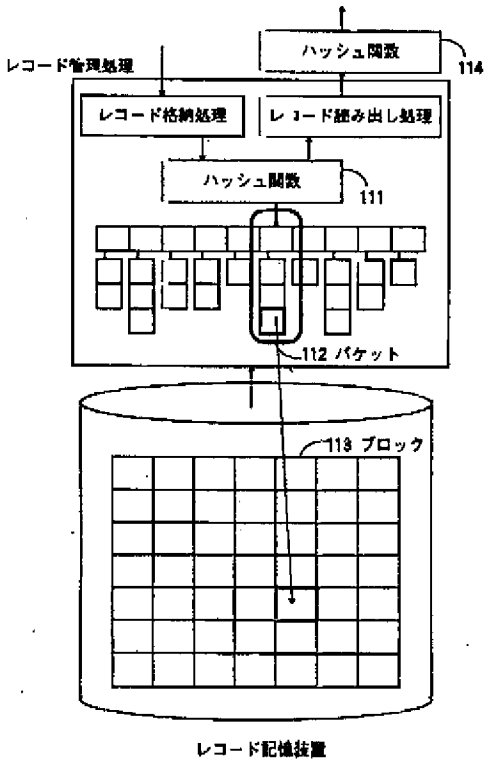
【図10】

図10



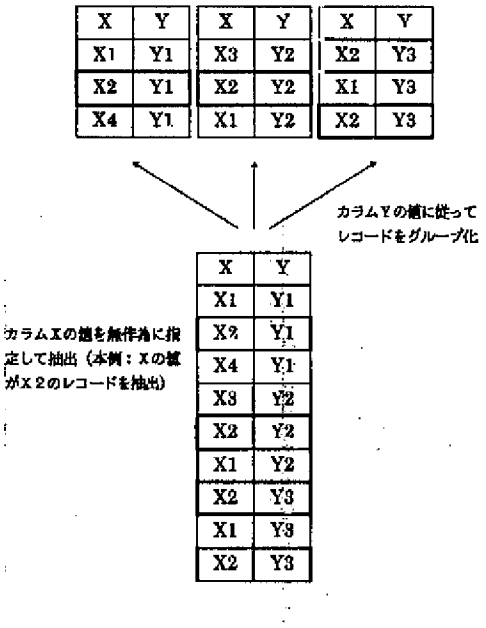
【図11】

図11



【図１７】

図１７



フロントページの続き

(72)発明者　高橋　ヨリ
神奈川県横浜市戸塚区戸塚町5030番地　株式会社日立製作所ソフトウェア開発本部内

(72)発明者　西澤　格
東京都国分寺市東恋ケ窪一丁目280番地　株式会社日立製作所中央研究所内